基于科技文献多重共现的数据模型理论与知识发现应用范例研究*

■ 庞弘燊
深圳大学图书馆 深圳 518060

摘要:［目的/意义］科技文献中各种特征项及其之间的关联是构成多种多样共现现象的基本单元,通过挖掘共现特征项之间的关联,共现分析可以从不同角度探测科学与技术活动规律的方方面面,为科研管理者和研究者等提供一个全方位、多角度观察科学发展的新视角。［方法/过程］通过对多重共现的基础理论研究,构建一套独特的多重共现数据模型基础理论体系,该理论体系包括:多重共现的定义、多重共现的研究范畴、用于多重共现的变量符号、多重共现的矩阵定义、多重共现的数据组织形式以及多重共现的延展系数计算公式与应用范畴。此外,基于多重共现的交叉图可视化方式,构建可用于分析3个或以上特征项共现关系的知识发现方法,包括共现关联强度、被引关联强度以及共现突发强度的分析方法。［结果/结论］通过该基础理论体系的构建,拓展共现现象的研究范围,为共现分析走向多角度、多维度的多重共现分析提供基础理论的支持。并通过实证研究,选取不同的多重共现应用案例,证明该方法可应用于研究领域、研究机构、机构间对比、研究学者等方面的分析,同时具有较好的分析效果。由于该方法体系具有分析角度多维化和分析方法多样化的特点,通过该方法的分析,除能够实现一重、二重共现等的分析效果外,还能揭示出比一般共现更为广泛和深入的知识内容。

关键词: 多重共现 多特征项共现 多源数据 数据模型 知识发现
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2019.09.007

1 共现的分析范畴

科技文献中的共现是指在论文、专利等文献中相同或不同类型特征项共同出现的现象,如多篇期刊论文之间共同出现的主题(关键词)、共同出现的合作作者、合作机构以及论文与关键词、机构与作者共同出现等,以及专利文献中共同出现的发明人、发明人与IPC分类号共同出现等都属于共现研究的范畴[1]。

共现分析是将各种信息载体中的共现信息定量化的分析方法,以心理学的邻近联系法则[2]和知识结构及映射原则为方法论基础。通过共现分析,人们可以发现研究对象之间的亲疏关系,挖掘隐含的或潜在的有用知识,并揭示研究对象所代表的学科或主体的结构与变化。在计算机技术的辅助下,共现分析在构建概念空间和本体实现语义检索、改进知识组织中文本分类效果、分析文献中知识内容关联、挖掘知识价值等方面彰显出独特的功能,正在成为支撑知识挖掘和知识服务的重要手段和工具。在知识表达中,能够体现信息的内容特征和外部特征不仅具有语义内涵而且是相互关联的,这些内容特征与外部特征共同构成了文本知识关联揭示和知识挖掘的基础[3]。

在文献计量数据中,共现现象并不是个例,而是大量存在于论文数据中的普遍现象。各种类型的特征项共现将离散的论文数据联结成一个有机的整体,可以从多个角度揭示科学活动规律,如期刊论文的关键词直接反映科学研究的主题及其细节、方法、技术,对关键词的共现现象进行分析可以用来考察科学在知识、

* 本文系教育部人文社会科学研究青年基金项目"基于科技文献多源数据融合及多特征项共现分析技术的科研动态识别监测方法研究"(项目编号:18YJC870015)和广东省哲学社会科学一般项目"学科领域创新演化路径的情报分析方法研究"(项目编号:GD18CTS03)研究成果之一。

作者简介: 庞弘燊(ORCID:0000-0002-5039-8817),副研究馆员,博士,E-mail:phs@szu.edu.cn。

收稿日期:2018-02-27 修回日期:2018-07-23 本文起止页码:61-72 本文责任编辑:易飞

方法、维度上的结构；专利发明人作为技术发明的主体，发明人共现研究是专利技术合作在个人层面的直接表征；论文和专利各自及相互之间的引用，作者/发明人之间、研究团体之间、机构之间乃至国家之间的引用是无形学院的有形标志，通过对这些引用与被引用现象的分析可以获知科学交流的模式、规律和特征。

由此看来，各种特征项及其之间的关联是构成多种多样共现现象的基本单元，通过挖掘共现特征项之间的关联，共现分析可以从不同角度探测科学（主要通过论文表征）与技术（主要通过专利表征）活动规律的方方面面，为科研管理者、科学研究者、技术发明者和技术应用者等提供一个全方位、多角度观察科学发展的新视角。

到目前为止，国内外许多研究学者已经对论文中特征项之间的共现分析方法和工具进行了多方面的研究。如 R. Fano[3] 在1956年首次提出文献耦合的概念和思路；H. Small[4] 提出了共被引分析概念；H. White 和 B. Griffith[5] 提出作者共被引分析方法；M. Callon 等[6] 首次提出了共词分析方法；中国学者郑华川等[7] 提出共篇分析；D. Zhao 等[8] 提出了作者参考文献耦合分析；刘志辉等[9] 提出作者关键词耦合分析方法；L. Yang 等[10] 应用了机构与关键词共现分析方法。而在共现分析的可视化工具上，共涉及10多种软件工具，包括科学计量学研究软件 Bibexcel、统计学软件 SPSS、引文网络可视化软件 CiteSpace、社会网络分析软件 Ucinet 和 Pajek，以及其他共现分析工具如 SCI-map（引文网络浏览）、Histcite（引用分析）、DIVA（文献耦合、合著分析）等，这些软件工具可对不同类型的共现分析进行可视化显示，有助于共现分析的有效解读和可视化展示。

但是，对于多特征项共现进行过相关揭示和分析的学者并不多见。其中有美国科学计量专家 Morris[11-12] 为了揭示两种特征项之间的关联，与其团队借助于两个共现矩阵相同特征项之间的关联，开发了交叉图和时间线技术并进行了应用研究，两种技术可以很好地弥补目前可视化技术不能揭示两种特征项关联的缺陷。冷伏海等[13] 认为目前共词分析研究主要关注二元词对共现的研究，对三元甚至多元词组的共现很少涉及，并在研究中提出基于位向量的三元共词分析算法和基于坐标图的三元共词结果分析方法。张自立等[14] 认为文献特征共现分析可以揭示文献的内容关联和特征项隐含的寓意，并基于2—模网络模型探讨不同特征共现的分析方法，有利于挖掘不同特征共现网络的深层次结构关系。L. Leydesdorff[15] 把“异质网络”的思想进一步扩展到了3—模网络，他把作者—期刊—关键词的特征项联系起来，通过不同类型节点在同一网络中的展现，不仅有利于分析同一类型节点间以及不同类型节点间的关系，而且也是研究网络更加真实的反映。

目前国内外对特征项共现的研究方法以及工具软件多集中在两个特征项之间共现的研究[16-20]，并且多是通过融合多种两个特征项共现的方法来揭示多特征项共现的关系[21]，而直接对3个或以上特征项之间的共现分析方法及可视化方式的研究并不多见，庞弘燊等人[22-24] 使用多重（多特征）共现的分析方法并开发相应的可视化图谱分析工具来对科研机构/科研领域进行分析，分析的视觉和范围大大扩大，能够发掘出比一般共现更深入和广泛的知识。可见，如果能够直接从3个或以上特征项共现的视觉出发，通过系统的知识发现方法研究来揭示3个或以上特征项之间的共现关系，在反映科学技术活动规律和科学知识领域方面可以增加多个分析角度和信息来源，有很大的知识挖掘和探索价值。此外如果能通过整合科技文献数据库的多源数据进一步挖掘共现特征项之间的关联，基于科技文献多源数据融合及多特征项共现分析技术的情报分析方法将有利于研发融合多源大数据的个性化价值发现方法，并通过研究多学科科技文献领域的数据融合和关联关系，发掘学科一般发展及其交叉发展的价值范式。

2 多重共现的定义与研究范畴

本论文把科技文献（包括论文、专利等文献）中单个特征项在多篇文献中的重复出现称作一重共现，两个特征项的共现称为二重共现，以此类推，3个或以上特征项共现都称作多重共现。因此，本论文把多重共现（multiple occurrence）定义为3个或以上相同类型或不同类型特征项共同出现的现象，如作者—关键词—发表期刊3个特征项同时在多篇论文中出现，发明人—IPC分类号—关键词、作者—引文作者—关键词—引文关键词等3个或以上特征项的共现都属于多重共现研究的范畴。

多重共现与二重共现相比，如作者—关键词—发表期刊的多重共现比作者—关键词、作者—发表期刊等的二重共现能够揭示更为深入的知识。分析作者—关键词—发表期刊的多重共现就相当于同时分析作

者—关键词、作者—发表期刊、关键词—发表期刊这3个二重共现现象及其之间的关系。如表1所示,可以看出多重共现现象对于揭示深度知识方面有着独特的优势,特别是在分析论文—专利多特征项交叉共现时,还能够进一步反映产学研的演化路径发展关系。图1则直观地显示多重共现与一般特征项共现分析对象的区别。图2显示出多源科技文献(论文、专利、专著等)在多重共现中可能会呈现的关联关系示例。

表1 不同特征项共现所能揭示的知识内容

特征项共现个数	例子	分析的视角	所能揭示的知识
一个特征项(一重共现)	作者	高产作者	高发文量的作者
	关键词	高频关键词	热门研究主题词
两个特征项(二重共现)	关键词—关键词	共词分析	关键词聚类揭示研究主题
	作者—关键词	作者与关键词关系分析	作者的研究领域
三个或以上特征项(多重共现)	作者—关键词—发表期刊	作者、关键词与发表期刊之间的关系分析	作者偏好在某期刊上所发表的主题类型、某期刊的固定作者群及主题研究领域与变化等
	作者—关键词—引文关键词	作者、关键词与引文关键词之间的关系分析	通过关键词聚类和引文关键词聚类共同反映作者的研究领域
论文—专利多特征项交叉共现(多重共现)	论文关键词—专利主题词—专利引文—论文引文	论文与专利文献在关键词和引文间的相互关系分析	反映基础研究—应用研究在关键技术节点上演化路径的变化情况
	论文作者—专利发明人—论文单位—专利权人	论文与专利文献在作者和机构间的相互关系分析	反映基础研究—应用研究在研发人员和研发机构之间的关联情况,有助于找出该领域的关键技术人员与重点研发和应用机构

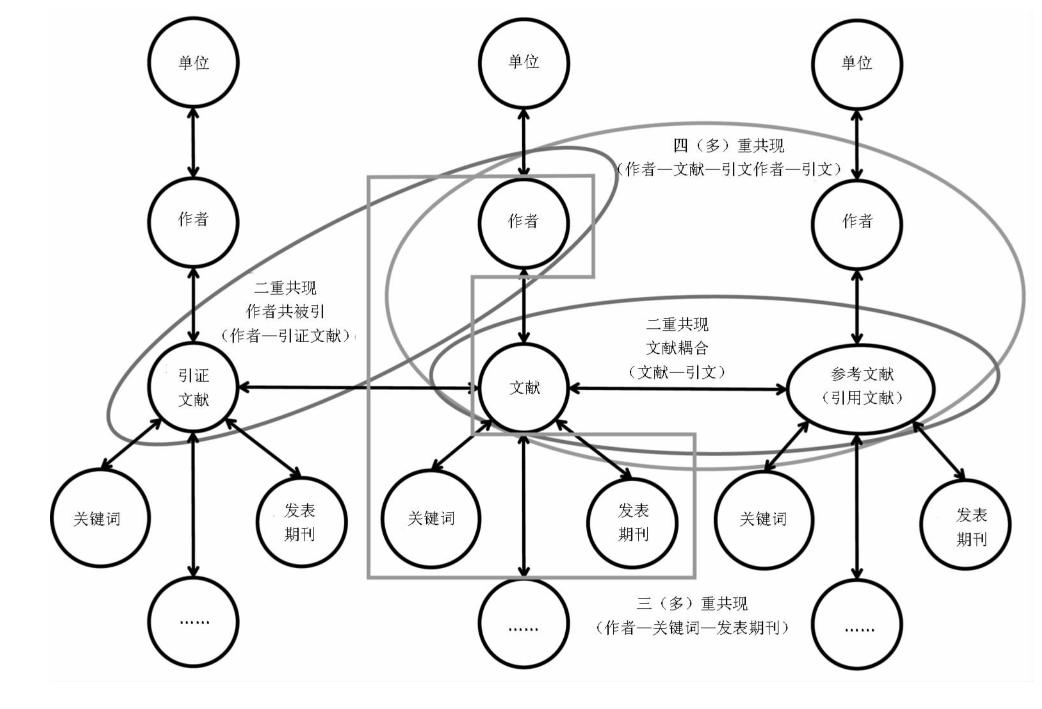


图1 多重共现与二重共现研究对象的区别(期刊论文)

3 多重共现特征项的变量符号

在S. Morris的博士论文中,使用图3(本论文进行了编译)形象地描绘出在期刊论文中各特征项之间的联系。箭头所示及箭头旁边的文字所述代表两个特征项之间的交互关系,如论文→关键词,代表了不同的关键词都可以在多篇论文中出现多次;而关键词→论文,则代表了每篇论文可以对应包含多个关键词。

并且S. Morris也用各特征项名称的缩写作为变量的名称来代表特征项,因此,本论文的研究当中也沿用了部分S. Morris用于定义期刊论文中不同特征项的变量符号,并对其在科技文献(包括论文和专利)中的应用进行了扩展(见表2)。

4 多重共现的矩阵定义与数据组织形式

在文献计量研究中,为了实现对共现现象的数据

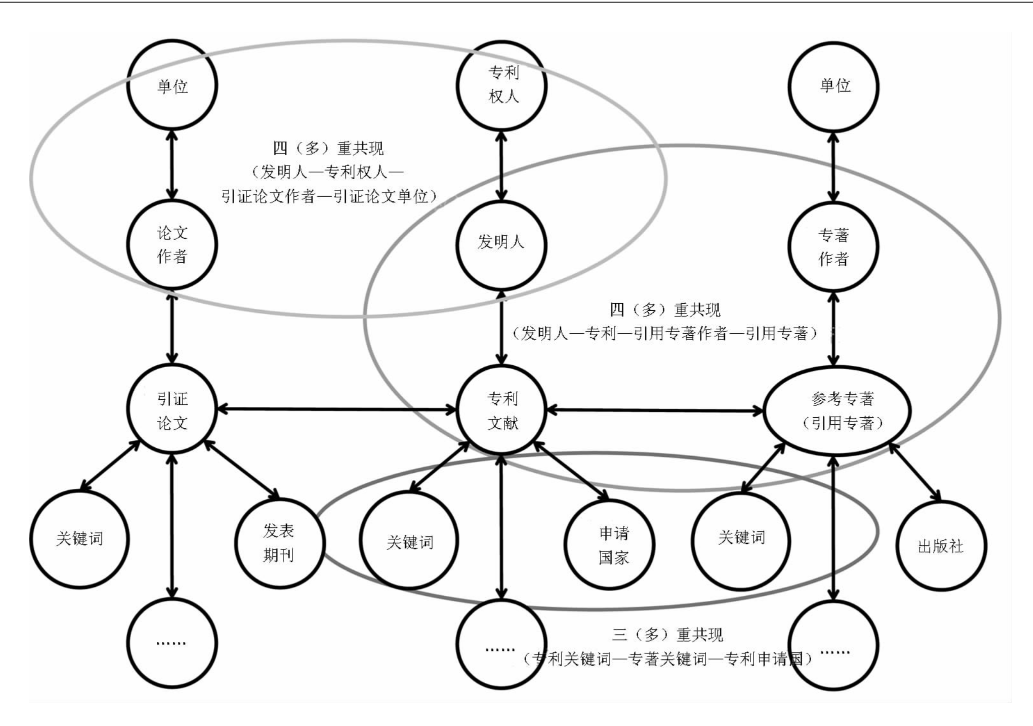


图2　多源科技文献（论文、专利、专著）间的多重共现关系示例

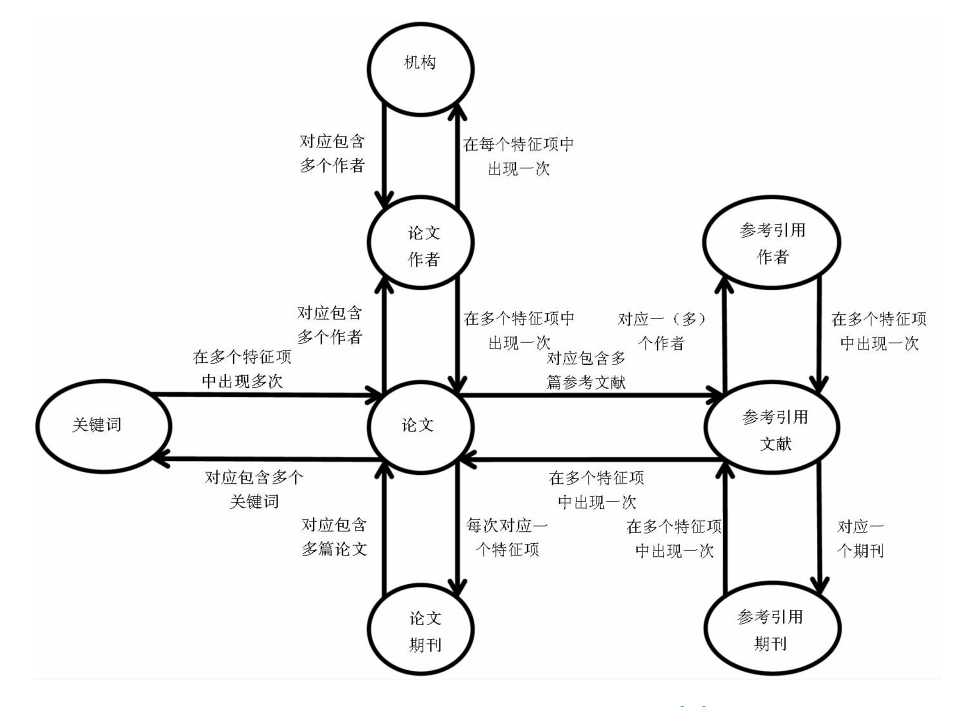


图3　S. Morris的期刊论文特征项关系[11]

挖掘，用定量分析方法来测度共现特征项之间的关联结构，首先要对数据进行数学处理，转换为各种共现矩阵，在此基础之上，运用数据挖掘以及各种可视化的分析方法找到隐含在矩阵中的数据关系。各种共现分析虽然在应用层面上揭示了不同的科学活动现象，但矩阵分析技术研究却大同小异[1]。

共引、共词及其他同种特征项共现矩阵在情报学领域应用极为广泛，在共现研究的早期，由于计算机存储技术、处理数据速度及基于矩阵的数据挖掘技术的限制，很多数据分析是基于同种特征项共现矩阵的，随着计算机技术日新月异的发展，可视化技术对多个矩阵转换的需求不断增加，研究者逐步认识到矩阵转换研究的重要性。荷兰莱顿大学的学者E. Englesman和A. van Raan发现原始的二值共现矩阵可以通过矩阵乘法转换为相应的对称共现矩阵[25]。美国科学计量专家S. Morris在博士论文中将各种共现矩阵之间的

数学转换关系作了系统和全面的研究[11]。

表 2　代表不同特征项的变量符号

变量符号	英文名	中文名
p	paper	论文
ap	paper author	论文作者
jp	paper journal	论文期刊
yp	paper year	发表论文的年份
ip	paper institution	发表论文的单位
kwp	paper keyword	论文关键词
r	reference	参考文献
ar	reference author	参考文献的作者
jr	reference journal	参考文献的期刊
yr	reference year	参考文献的年份
ir	reference institution	参考文献的单位
kwr	reference keyword	参考文献的关键词
pat	patent	专利
pi	patent inventor	专利发明人
pa	patent assignee	专利权人
py	patent application year	专利申请年
pc	patent applicant country	专利申请国
pkw	patent keyword	专利关键词
pr	reference patent	参考专利
……	……	……

在人工智能、工程、物理、化学、计算机科学等领域中,图结构被广泛应用,是一种很好的数据关系表现方式,图 4 明晰地表达了两个特征项(论文与参考文献)之间的关系,展示了共现现象背后特征项之间的关联结构。然而,要对图结构所表示的各种复杂关系进行描述,将图的关联结构存储在计算机中,必须将图结构转换为结构化的数据以便处理。在文献计量研究中,引入矩阵描述的方法来表达共现的关系网络。

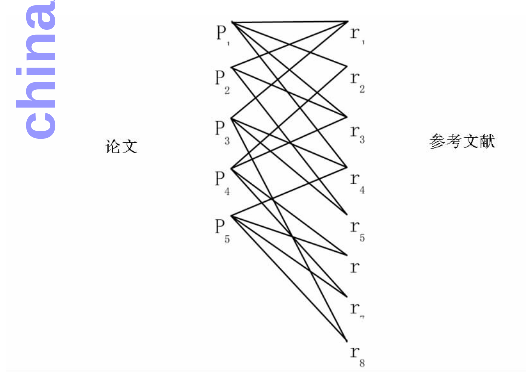


图 4　论文与参考文献之间关系的图结构[11]

首先给出矩阵的通用定义:由 $m*n$ 个数 $a_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$ 排成的 m 行 n 列的表:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{nm} \end{bmatrix}$$

称为一个 m 行 n 列的矩阵或 $m*n$ 矩阵,简记为 $A=(a_{ij})m*n$。数 a_{ij} 称为矩阵 A 的第 i 行第 j 列或(i,j)元素,i 称为元素 a_{ij} 的行标,j 称为元素 a_{ij} 的列标。特殊地,$n*n$ 矩阵也称为 n 阶方阵。

在社会网络分析中,不对称矩阵的列和行分别代表行动者(actor)和指标,对于对称的正方阵,行与列代表完全相同的行动者;在文献计量研究中,对于共现现象的矩阵描述,赋予共现矩阵的行与列特定的含义:行与列分别代表共同出现的特征项。矩阵中的元素代表行与列对应特征项之间是否相关或者关系的强弱。

在 S. Morris[11] 的博士论文当中,其对两个特征项共现现象的矩阵定义为:

$$O_{ij}[x_1;\ x_2]=\begin{cases} n & \text{特征项 i 与特征项 j 共同出现的频次为 n} \\ 0 & \text{特征项 i 与特征项 j 没有共同出现} \end{cases}$$

x_1、x_2 代表两种不同的特征项(如关键词、作者、发表期刊等),i、j 分别为 x_1、x_2 的具体对象。

其对应的图结构见图 5。

本文在 S. Morris 研究的基础上,将其对二重共现分析的研究理论扩展到多重共现领域,包括把 S. Morris 对共现的矩阵定义扩展到多重共现领域,同时在共现数据组织形式上也从二维的矩阵形式扩展到多元组的表示形式,以适用于多重共现的分析。

本文把 S. Morris 的矩阵定义扩展到多重共现领域特征项的共现关系,并定义:

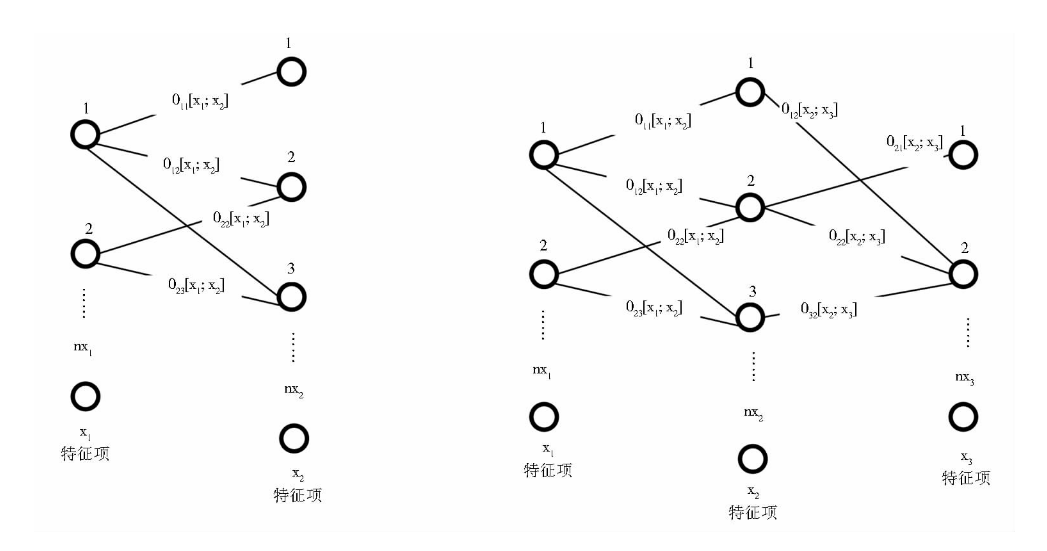
$$O_{ijk\cdots}[x_1;\ x_2;\ x_3\cdots]=\begin{cases} n & \text{特征项 i 与特征项 j、k 等共同出现的频次为 n} \\ 0 & \text{特征项 i、j、k 等没有共同出现} \end{cases}$$

其中,该多维矩阵定义所代表的图结构如图 6 所示,相同线型的连线代表该几个特征项之间共同出现的频率,如 $o_{111}[x_1;x_2;x_3]$ 代表了特征项集合 x_1 中序号为 1 的特征项与 x_2、x_3 中序号为 1 的特征项所共同出现的频次。

在数据组织形式上,S. Morris 使用的是传统二维矩阵来表示两个特征项之间的共现关系[1]:

$$O[x_1;\ x_2]=\begin{bmatrix} O_{11} & O_{12} & \cdots & O_{1nx} \\ O_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ O_{nx_11} & \cdots & \cdots & O_{nx_1nx_2} \end{bmatrix}$$

由于多重共现领域涉及到的是 3 个或以上特征项的关联关系,传统的二维矩阵的数据组织形式已不能适用于多重共现分析的要求,因此本文中通过使用多元组 $R(x_1,x_2,x_3\cdots,value)$ 来表示多维数据信息,用于分析多重共现特征项之间的关系。定义 $value_{ijk\cdots}$ 为 x_1 中特征项 i 与 x_2 中特征项 j、x_3 中特征项 k 等共同出现的频次,即 $value_{ijk\cdots}=O_{ijk\cdots}[x_1;\ x_2;\ x_3\cdots]$。通过从二维矩阵扩展到多元组的数据表示形式(见图 7),以适用于多重共现的数据组织和多特征项的共现分析。



注:序号 1 至 nx_m 代表特征项,该特征项可以代表论文或专利的关键词、作者(发明人)、等类型;特征项之间的连线代表两个特征项共现,其共现频次用数值 $O_{ij}[x_m; x_n]$ 来标识

图 5　Morris 的特征项组对图结构[11]

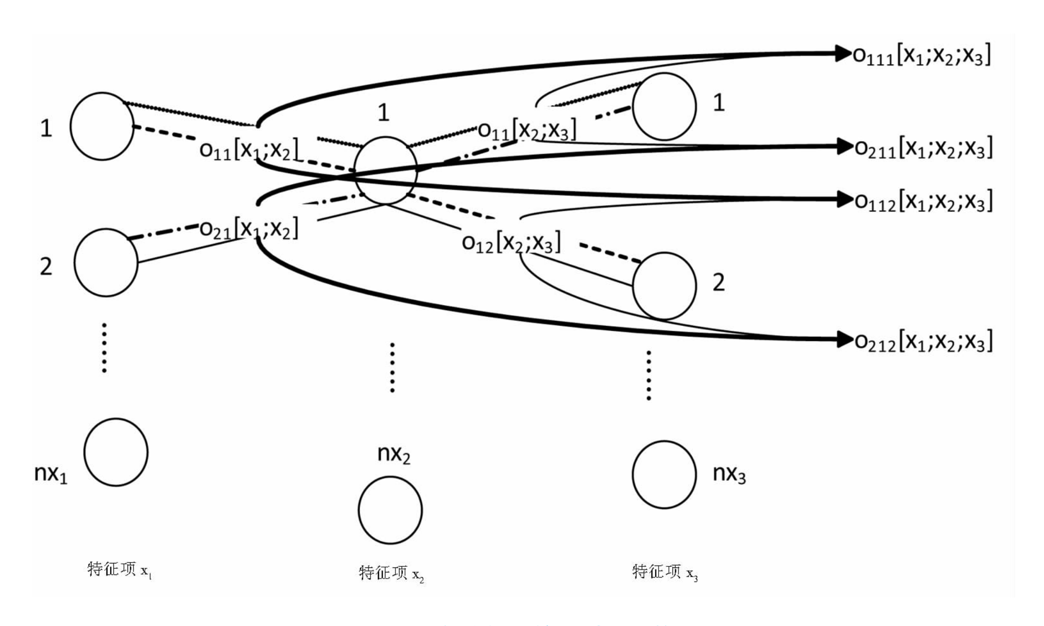


图 6　多特征项的组对图结构

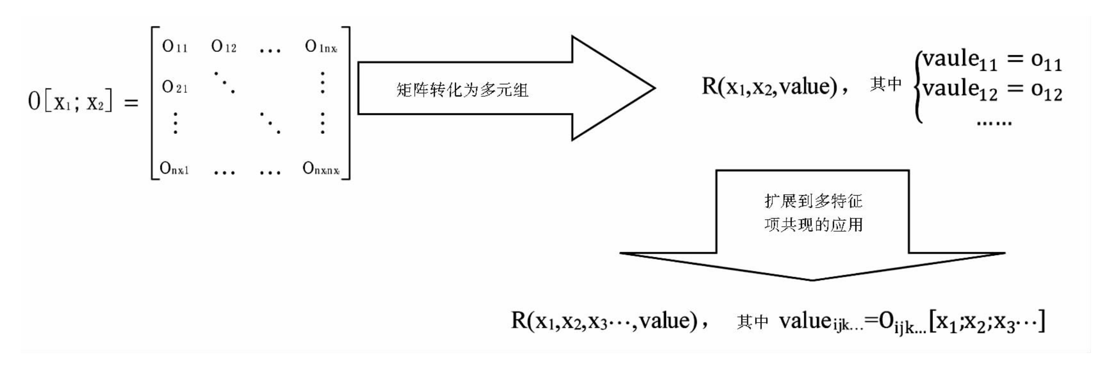


图 7　多特征项共现分析中数据组织形式的变化

庞弘燊. 基于科技文献多重共现的数据模型理论与知识发现应用范例研究[J]. 图书情报工作,2019,63(9):61 – 72.

在上述的矩阵定义中,使用了共现频数作为共现矩阵元素的值。本文还使用了二值法对其进行定义,在二值矩阵中,所有元素的取值均为 0 或 1。因此本文在 $O_{ij}[x_1;x_2]$、$O_{ijk\cdots}[x_1;x_2;x_3\cdots]$定义的基础上,对 $O'_{ij}[x_1;x_2]$和 $O'_{ijk\cdots}[x_1;x_2;x_3\cdots]$二值矩阵定义如下:

$$\text{对 } O'_{ij}[x_1;x_2]=\begin{cases}n & \text{特征项 i 与特征项 j 共同出现的频次为 1 次或以上}\\ 0 & \text{特征项 i 与特征项 j 没有共同出现}\end{cases}$$

$$O'_{ijk\cdots}[x_1;x_2;x_3\cdots]=\begin{cases}1 & \text{特征项 i 与特征项 j,k 等共同出现的频次为 1 次或以上}\\ 0 & \text{特征项 i,j,k 等没有共同出现}\end{cases}$$

以下举例说明 $O_{ijk\cdots}[x_1;x_2;x_3\cdots]$和 $O'_{ijk\cdots}[x_1;x_2;x_3\cdots]$所代表的共现意义,假设有专利文献数据集 D1 如表 3 所示:

表 3　多重共现矩阵定义示例数据集

数据集 D1	发明人(pi)	申请国家(pc)	申请年份(py)	关键词(pkw)
专利 1	发明人 1,发明人 2	国家 1	年份 1	关键词 1,关键词 2,关键词 3
专利 2	发明人 2,发明人 3	国家 1	年份 2	关键词 4,关键词 5
专利 3	发明人 1	国家 2	年份 2	关键词 3,关键词 4

对于整个数据集 D1 来说:

O[发明人 1;关键词 3] = 2,代表在数据集 D1 中,发明人 1 与关键词 3 共现 2 次。

O′[发明人 1;关键词 3] = 1,代表在数据集 D1 中,发明人 1 与关键词 3 存在共现(频次在 1 次或以上)。

O[发明人 1; 国家 2;关键词 3] = 1,代表在数据集 D1 中,发明人 1 与国家 2、关键词 3 共现 1 次。

O′[发明人 1; 国家 2;关键词 5] = 0,代表在数据集 D1 中,发明人 1 与国家 2、关键词 5 没有共同出现过。

5　多重共现的延展系数

基于以上多重共现矩阵关系的定义可用于计算多重共现特征项的延展系数 E_{Xn}和 E'_{Xn}。

定义 m_i、……、m_j、m_k 分别为特征项 x_1、……、x_{n-1}、x_n 包含的所有不同的对象数,则有公式(1) – 公式(4):

$$E_{Xn}(i,\cdots,j)=\frac{\sum_{k=1}^{mk}O_{i,\ldots\ldots,jk}[x_1;\cdots;x_n]}{O_{i\cdots j}[x_1;\cdots;x_{n-1}]}\qquad \text{式(1)}$$

$$E_{Xn}(x_1;\cdots;x_{n-1})=\frac{\sum_{i=1}^{mi}\cdots\sum_{j=1}^{mj}\sum_{k=1}^{mk}O_{i\cdots jk}[x_1;\cdots;x_n]}{\sum_{j=1}^{mi}\cdots\sum_{j=1}^{mj}O_{i\cdots j}[x_1;\cdots;x_{n-1}]}\qquad \text{式(2)}$$

$$E'_{Xn}(i,\cdots,j)=\frac{\sum_{k=1}^{mk}O'_{i\cdots jk}[x_1;\cdots;x_n]}{O'_{i\cdots j}[x_1;\cdots;x_{n-1}]}\qquad \text{式(3)}$$

$$E'_{Xn}(x_1;\cdots;x_{n-1})=\frac{\sum_{i=1}^{mi}\cdots\sum_{j=1}^{mj}\sum_{k=1}^{mk}O'_{i\cdots jk}[x_1;\cdots;x_n]}{\sum_{i=1}^{mi}\cdots\sum_{j=1}^{mj}O'_{i\cdots j}[x_1;\cdots;x_{n-1}]}\qquad \text{式(4)}$$

延展系数 E_{Xn}和 E'_{Xn}可应用的范畴为:

E_{Xn}:用于分析某特征项在每一篇科技文献中的平均数量分布状况,如每篇期刊论文平均采用多少个关键词,某年申请的专利平均有多少个发明人,在某年某期刊上论文的平均作者数、平均关键词数的多少等。

E'_{Xn}:用于分析某特征项在整个数据集内种类的分布状况,如某作者在多少种期刊上或多少年内发表过论文,某发明人在某年内申请了多少种类型的专利,某期刊在多少年内刊载过某作者的论文等。

以下举例说明延展系数 E_{Xn}和 E'_{Xn}所能揭示的意义,假设有期刊论文数据集 D2,如表 4 所示:

表 4　多重共现延展系数示例数据集

数据集 D2	作者(ap)	发表期刊(jp)	发表年份(yp)	关键词(kwp)
论文 1	作者 1,作者 2	期刊 1	年份 1	关键词 1,关键词 2,关键词 3
论文 2	作者 2,作者 3	期刊 2	年份 1	关键词 4,关键词 5
论文 3	作者 1	期刊 2	年份 2	关键词 1,关键词 2

对于论文 1 来说:

E_{kwp}(ap) = 3,代表论文 1 的每个作者都用了三个关键词。

E_{kwp}(ap;jp) = 3,代表论文 1 的每个作者在每种发表论文的期刊上都用了三个关键词。

对于整个数据集 D2 来说:

E_{kwp}(作者 1) = 2.5,代表在数据集 D2 中,作者 1 发表的每篇论文平均用了 2.5 个关键词。

E_{kwp}(作者 1,期刊 1) = 3,代表在数据集 D2 中,作者 1 在期刊 1 上发表的每篇论文平均用了 3 个关键词。

E'_{jp}(作者 1) = 2,代表在数据集 D2 中,作者 1 在两

种不同的期刊上发表过论文。

E'_{yp}(作者 2) = 1,代表在数据集 D2 中,作者 2 只在一年里发表过论文。

E'_{jp}(作者 1,年份 1) = 2,代表在数据集 D2 中,作者 1 在年份 1 内只在一种期刊上发表过论文。

E_{kwp}(ap) = 2.4,代表在数据集 D2 中,每个作者发表的每篇论文平均用了 2.4 个关键词。

E_{kwp}(ap,jp) = 2.4,代表在数据集 D2 中,每个作者在每个期刊上发表的每篇论文平均用了 2.4 个关键词。

E'_{jp}(ap) = 1.67,代表在数据集 D2 中,平均每个作者在 1.67 种不同的期刊上发表过论文。

E'_{yp}(ap) = 1.33,代表在数据集 D2 中,平均每个作者在 1.33 个不同的年份内发表过论文。

E'_{jp}(ap,yp) = 1.25,代表在数据集 D2 中,平均每个作者—年份组合在 1.25 种不同的期刊上发表过论文,即在活跃年(有论文发表的年份)内的活跃作者(有论文发表的作者)平均在 1.25 种期刊上发表了论文。

6 多重共现的知识发现方法体系设计

本文把知识发现的概念、模式、一般过程与多重共现的分析过程结合起来,在设计多重共现知识发现方法的分析过程中也遵循以下一般的知识发现分析步骤:多源科技文献数据搜集与清理→数据处理(使用矩阵转换技术、降维技术、聚类分析等)→生成多重共现交叉图→分析多重共现交叉图特点→汇总知识发现结论。本文设计的多重共现知识发现的方法体系如图 8 所示,包括共现关联强度的分析、被引关联强度的分析、共现突发强度的分析 3 个方面的内容。

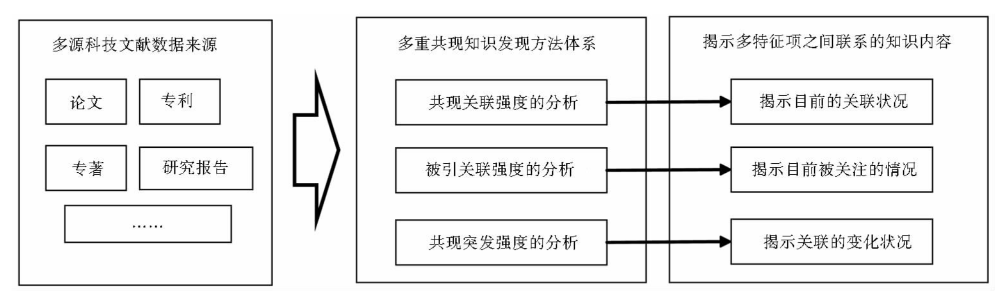


图 8 多重共现的知识发现方法体系

共现关联强度分析是指通过对多个特征项之间共现频次大小的分析,揭示其潜在的共现关联状况;被引关联强度分析是指通过对多个特征项之间共同被引频次大小的分析,来揭示其被关注的情况;而共现突发强度的分析是指通过对多个特征项共现突发权值的分析,来揭示其变化状况及突发的热点内容。

通过该方法体系的构建可以完善多重共现的知识发现方法,从多源数据来源并从多个角度来揭示多特征项之间的关联知识,包括对单个特征项的聚类或频次分析、两个特征项之间的关联关系、乃至多个特征项之间的关联关系。因此,多重共现知识发现方法的设计除了可分析三重或以上的多重共现之外,还同时涵盖了一重、二重共现的分析。

在该方法体系下,多重共现的共现关联强度、被引关联强度以及共现突发强度的分析方法以及具体的分析流程和交叉图的可视化方式也不尽一致。该套多重共现的知识发现方法可以分析 3 个或以上论文特征项的关系,但是由于在涉及到 3 个以上特征项共现的时候,其共现的频次大多较低,数据离散程度较高,并不利于关联强度和突发强度的知识揭示。因此在下面的多重共现知识发现方法应用范例中,将主要基于 3 个特征项的共现(三重共现)作为多重共现的研究样例,而 3 个以上特征项共现的分析方法亦可依照此分析方法来作进一步的类推。此外该方法体系的数据来源可以来源于论文、专利、专著等科技文献,同时针对不同类科技文献间的多重共现组合所分析的角度以及分析意义都有着不少的差异,因此在实证分析中还需要根据研究目的和分析需求,选取科技文献的类型及其特征项组合来进一步分析和研究。

7 三重共现应用范例研究

多重共现的知识发现方法可根据具体研究目的,研究内容选定分析方法、多源科技文献数据集合或特征项的组合来进行分析。而且不同类型文献来源的多重共现项的组合,会依研究目的、分析方法和数据集的不同,而揭示出不同方面的知识内容。

从分析效果上看,在多重共现的知识发现方法当中,由于是基于多重共现交叉图的分析,因此通过多重

共现的知识发现方法就能够基于一个三重共现交叉图来同时实现一重、二重、三重共现的分析效果,除了提高分析效率之外,还可以从多个角度揭示出更为广泛和深入的知识内容,可见该知识发现方法具有一定的可行性,并比原来的一重、二重共现的可视化分析效果更好。

该套知识发现方法体系应用范围较为广泛,可以对研究领域、研究机构、机构间对比、研究学者等多个方面进行分析,并且可以依据分析的目的,选取该套方法体系中的一个或多个分析方法进行组合分析,此外如果结合论文和专利等多源科技文献数据进行分析,还可以进行产学研的创新演化路径分析等。图 9 – 图 11 所示的多组多重共现可视化图是对多重共现的知识发现方法体系进行了应用范例的实证研究,在实证中只选用了期刊论文的单源数据来源作为可视化示例。

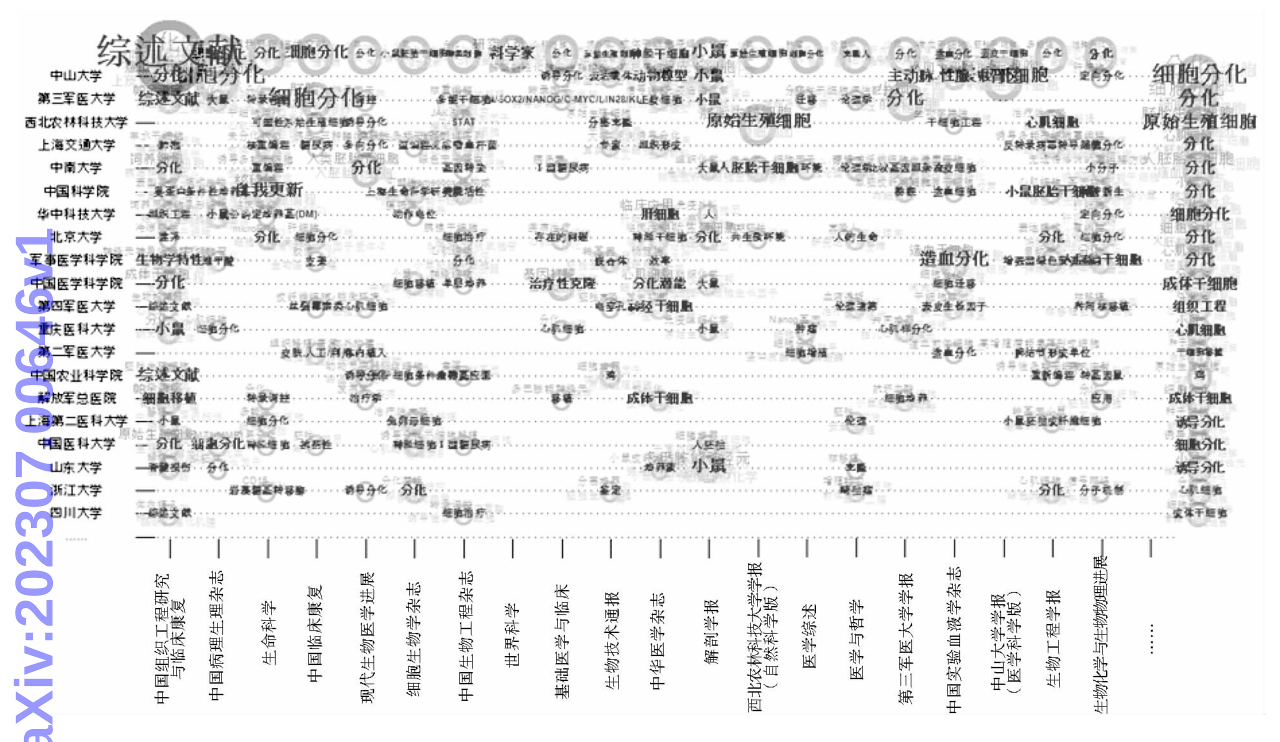


图 9　机构—期刊—关键词三重共现交叉图(研究领域:胚胎干细胞)

图 9 是针对研究领域分析的机构—期刊—关键词三重共现交叉图(研究领域:胚胎干细胞),其可用于分析在胚胎干细胞研究领域有哪些主要研究机构,并且分析其在主流发表期刊中的研究主题分布等等。从图的上下两侧区域可以看出,胚胎干细胞研究领域主流发表期刊载文量较多的期刊按由多到少顺序依次自左向右排列,胚胎干细胞领域载文量居前的期刊有《中国组织工程研究与临床康复》《中国病理生理杂志》《生命科学》等。各期刊的载文主题不尽相同,比如《中国组织工程研究与临床康复》刊载了较多关于胚胎干细胞的"综述文献",《解剖学报》刊载的主题以"小鼠"为主等。从图的中间区域可以看出各主要研究机构在主流发表期刊中的发文主题分布情况,比如中山大学在《中国病理生理杂志》上主要发表关于"细胞分化"与"造血干细胞"研究的论文,在《中山大学学报(医学科学版)》上发表了较多关于"表皮干细胞"的研究论文,而第三军医大学在《中国临床康复》与《第三军医大学学报》上发表了较多以"细胞分化"为研究主题的论文。从以上对胚胎干细胞研究领域的分析可以看出,基于多重共现的交叉图可视化技术以及知识发现分析方法能够较好地揭示出该研究领域当中多个特征项的关联关系,相关机构可以此作为参考,跟踪该领域的研究情况以及发展趋势。

图 10 是针对研究机构进行分析的年份—发表期刊—作者三重共现被引关联强度交叉图(研究机构:中国科学院文献情报中心),可以用于观测和计量研究机构中被引频次较高的作者及其在发表期刊、年份之间的被引分布情况与被引发展趋势,并发现机构中的高被引量作者、年份、发表期刊等。

图 11 是针对研究机构进行分析的多重共现突发关联强度交叉图(研究机构:中国科学院文献情报中心),在三重共现的突发强度分析中,通过观测特征项

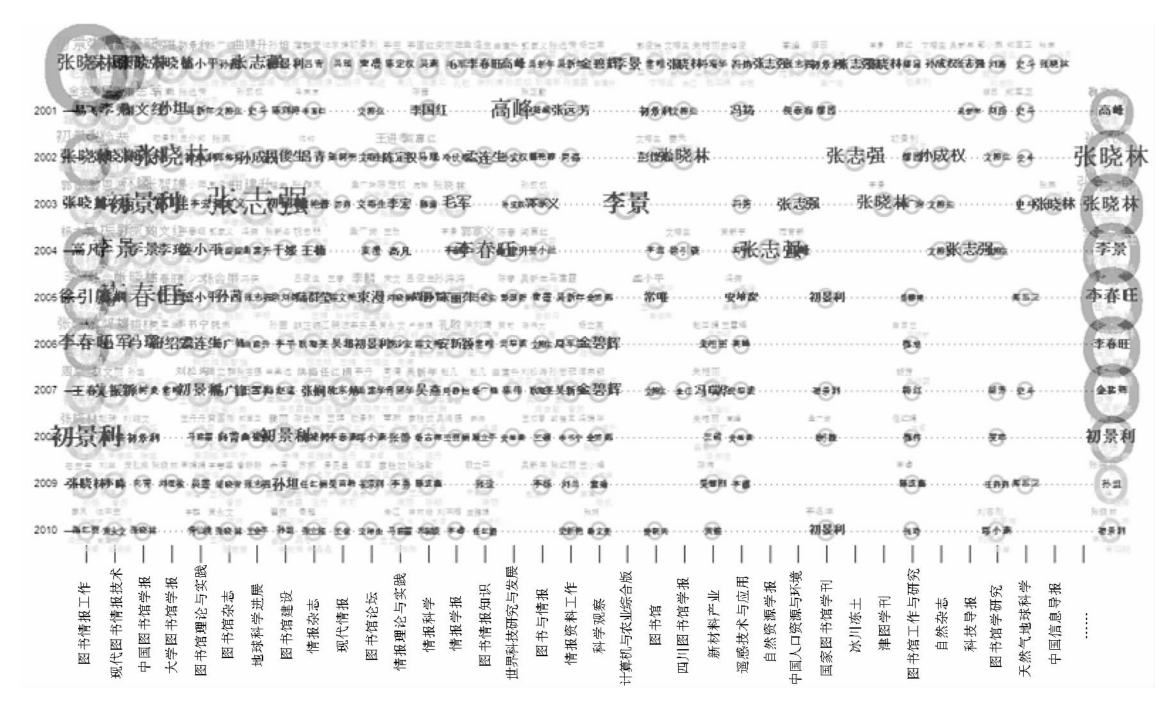


图 10 年份—发表期刊—作者三重共现被引关联强度交叉图(研究机构:中国科学院文献情报中心)

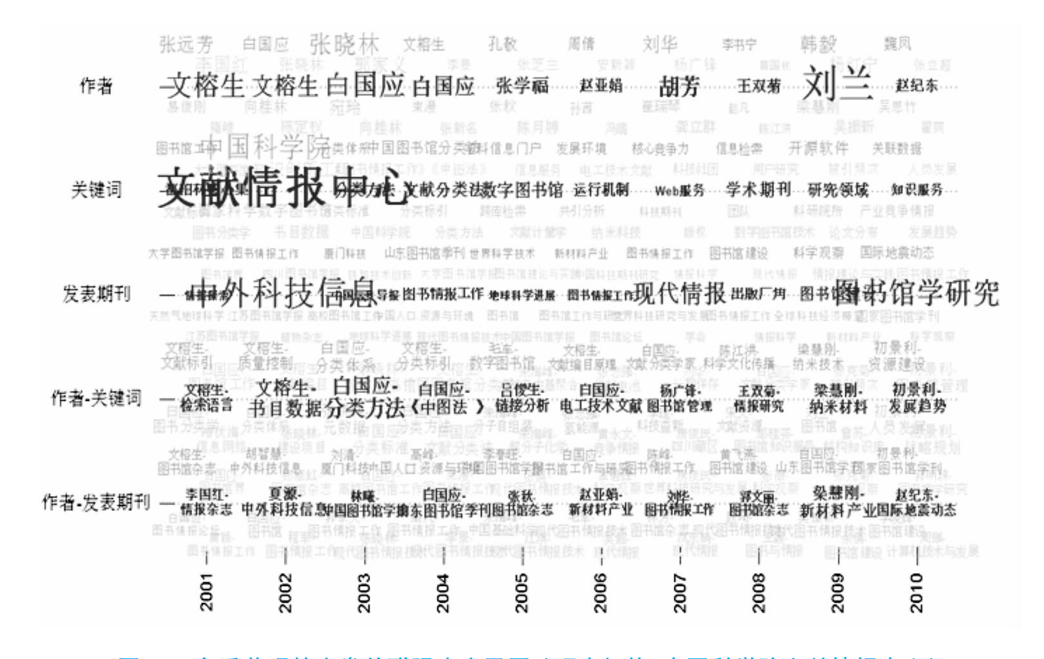


图 11 多重共现的突发关联强度交叉图(研究机构:中国科学院文献情报中心)

组合在某段时间内数据量的突发情况,可以发现其特征项组合的突发特征与热点研究内容,还可依此分析其突发原因和突发趋势。从作者突发强度看,在 2001－2010 年时间段的早期,以文榕生、白国应等为代表的科研人员发表的文章数增长较多;而在此时间段的后期,出现了较多中国科学院文献情报中心在读研究生的名字,表明这段时间内中国科学院文献情报中心的在读研究生发表文章的突发权值较高,是研究的新秀力量。在作者—关键词特征项的组合上,2001－2010 年间早期以文榕生、白国应为代表的研究分类学的关键词组合较多,后期热点慢慢转移并分化,形成了百花齐放的局面,可以看到不同作者与多样化研究主题的特征项组合迅速增长。

8 结语

本文对多重共现的相关概念进行了概述,界定了多重共现的定义和研究范畴,明晰了多重共现中特征项的变量符号。基于 S. Morris 原有共现研究的基础,

本文对多重共现的矩阵定义、数据组织形式以及延展系数的计算方式进行了研究。通过对多重共现的基础理论研究,本文构建了一套独特的多重共现基础理论体系,该理论体系包括:多重共现的定义、多重共现的研究范畴、用于多重共现的变量符号、多重共现的矩阵定义、多重共现的数据组织形式以及多重共现的延展系数计算公式与应用范畴。通过该基础理论体系的构建,拓展共现现象的研究范围,为共现分析走向多角度、多维度的多重共现分析提供了基础理论的支持。

而多重共现的可视化方式,除了多重共现交叉图(基于 S. Morris 交叉图的改进)的可视化方式外,其实还可以用基于社会网络分析方法的多模网络图显示,但是在实证研究中发现通过多重共现交叉图技术能够在一张图内同时显现出多模网络图中的 4 种共现关系,即在一个三重共现交叉图当中同时显现出 3 个特征项之间的 3 个二重共现关系(2 模网络图)以及一个三重共现关系(3 模网络图),在进行多重共现知识发现的分析过程中更为直观和便捷。并且交叉图技术相比多模网络图来看,在显示效果和数据方式上略胜一筹,因此在本研究当中,采取多重共现的交叉图技术作为可视化方式来进行多重共现的知识发现应用范例研究。

此外本文基于多重共现的交叉图可视化方式,构建了可用于分析 3 个或以上特征项共现关系的知识发现方法,包括共现关联强度、被引关联强度以及共现突发强度的分析方法。并通过实证研究,选取了不同的多重共现应用案例,证明该方法可应用于研究领域、研究机构、机构间对比、研究学者等方面的分析,同时具有较好的分析效果。由于该方法体系具有分析角度多维化和分析方法多样化的特点,因此通过该方法的分析,除了能够实现一重、二重共现等的分析效果外,还能揭示出比一般共现更为广泛和深入的知识内容。

在多重共现的研究中,目前已针对各类特征项组合的多重共现(如机构—期刊—关键词、年份—关键词—机构、年份—关键词—期刊、作者—期刊—关键词、年份—期刊—作者等)进行了分析,针对其他不同特征项组合的多重共现(如作者—年份—参考文献、作者—引证作者—引证年份等)的知识发现效果,以及针对多源科技文献数据(论文、专利和专著等)等的多重共现分析效果,还有待进一步研究和证明。结合多重共现的数据模型理论和交叉图知识发现的可视化应用案例方式,可根据具体特征项的组合以及具体科技文献类型组合来研究更有针对性的数据挖掘算法以增强和展示深度知识发现的效果。将来的研究将继续引入相关知识发现的理论方法,如数据挖掘与知识发现、决策树、关联规则、神经网络等技术,以发掘更多更深入的多特征项之间、多源科技文献之间共现的一般规律与特殊规律,并可以对其知识发现效果继续进行深入研究,对不同特征项组合和不同类型科技文献组合的分析效果进行归纳和总结。此外基于多重共现的知识发现方法还可以针对不同的科学领域,如针对其他自然科学、社会科学等不同领域进行知识发现,进行更多的实证研究,以进一步验证多重共现知识发现方法的可行性和适用范畴。

参考文献:

[1] 杨立英. 科技论文共现理论与应用[D]. 北京:中国科学院文献情报中心, 2007.

[2] 王曰芬,宋爽,苗露. 共现分析在知识服务中的应用研究[J]. 现代图书情报技术, 2006(4):29－34.

[3] FANO R, Information theory and the retrieval of recorded information[M]//Documentation in Action. New York: Reinhold Publ. Co., 1956:238－244.

[4] SMALL H. Maero-level changes in the strueture of co-eitation clusters:1983－1989[J]. Scientometries, 1993, 26(1):5－20.

[5] WHITE H, GRIFFITH B. Author co-citation: a literature measure of intelleetual structure[J]. Journal of the American Society for Information Scienee,1981, 32(3):163－169.

[6] CALLON M, LAW J, RIP A. Mapping the dynamics of science and technology: sociology of science in the real world[M]. New York: Sheridan House,1986.

[7] 郑华川,崔雷. 胃癌前病变低频被引论文的共词和共篇聚类分析[J]. 中华医学图书情报杂志, 2002, 11(3):1－3.

[8] ZHAO D, ANDREAS S. Evolution of research activities and intellectual influences in information science 1996－2005: introducing author bibliographic-coupling analysis[J]. Journal of the American Society for Information Science and Technology, 2008, 59(13): 2070－2086.

[9] 刘志辉,张志强. 作者关键词耦合分析方法及实证研究[J]. 情报学报, 2010, 29(2): 268－275.

[10] YANG L, MORRIS S, BARDEN E. Mapping institutions and their weak ties in a specialty: a case study of cystic fibrosis body composition research[J]. Scientometrics,2009(2):421－434.

[11] MORRIS S. Unified mathematical treatment of complex cascaded bipartite networks: the case of collections of journal papers [D]. Oklahoma: Oklahoma State University, 2005.

[12] MORRIS S, DEYONG C, WU Z, et al. DIVA: a visualization system for exploring document databases for technology forecasting [J]. Computers & industrial engineering, 2002, 43 (4): 841－862.

[13] 冷伏海,王林,李勇. 基于文献关键词的三元共词分析方法——

以知识发现领域为例[J]. 情报学报,2011(10):1072 - 1077.
[14] 张自立,张紫琼,李向阳. 基于 2 - 模网络的科研单位和关键词共现分析方法[J]. 情报学报,2011(12):1249 - 1260.
[15] LEYDESDORFF L. What can heterogeneity add to the scientometric map? steps towards algorithmic historiography[EB/OL]. [2018 - 01 - 30]. http://arxiv. org/abs/1002.0532.
[16] LEYDESDORFF L, VAUGHAN L. Co-occurrence matrices and their applications in information science: extending ACA to the web environment[J]. Journal of the American Society for Information Science and Technology,2006,56(12): 1616 - 1628.
[17] CHEN C, IBEKWE-SANJUAN, F, HOU J. The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2010,61(7):1386 - 1409.
[18] 张婷. 科学传播研究的可视化分析[D]. 大连:大连理工大学,2009.
[19] 刘则渊,陈悦,侯海燕,等. 科学知识图谱方法与应用[M]. 北京:人民出版社, 2008.
[20] 冯璐,冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报,2006(2):88 - 92.
[21] 胡琼芳,曾建勋. 基于多共现的文献相关度判定研究[J]. 情报理论与实践, 2010, 33(8):77 - 80.
[22] PANG H. A knowledge discovery method based on analysis of multiple co-occurrence relationships in collections of journal papers [J]. Chinese journal of library and information science,2012,5(4):9 - 20.
[23] 庞弘燊,方曙,范炜,等. 基于多重共现的机构科研状况分析方法研究——以中科院国家科学图书馆为例[J]. 情报学报,2012,31(11):1140 - 1152.
[24] 庞弘燊. 基于科技论文多特征项共现突发强度分析方法的算法实现与可视化图谱研究[J]. 图书情报工作,2015,59(24):115 - 122.
[25] ENGLESMAN E, VAN RAAN A. Mapping of technology: a first exploration of knowledge diffusion amongst fields of technology [R]. Bangalore:CWTS report, 1991.

Research on Data Model Theory and Knowledge Discovery Application Based on Multiple Occurrence of Scientific Literature

Pang Hongshen
Library, Shenzhen University, Shenzhen 518060

Abstract: [**Purpose/significance**] Various entities and their associations are the basic units that constitute a variety of occurrence phenomena in scientific literature. By mining the associations between occurrence entities, occurrence analysis can detect all aspects of the laws of scientific activities from different angles for scientific research management and researchers. It will provide a new perspective on the development of science from all angles and perspectives. [**Method/process**] By studying the basic theory of multiple occurrence, this paper constructs a set of unique basic theoretical system of multiple occurrence data model. The theoretical system includes definition of multiple occurrence, multiple occurrence research category, multiple occurrence variable symbols, multiple occurrence matrix definitions, multiple occurrence data organization forms, etc. In addition, based on the multiple occurrence cross-graph visualization method, this paper constructs a knowledge discovery method that can be used to analyze the occurrence relationship of three or more characteristic items, including the occurrence relevance strength, cited relevance strength and occurrence burst strength method. [**Result/conclusion**] Through the construction of this basic theoretical system, the research scope of occurrence phenomena is expanded, which provides the basic theory support for occurrence analysis to multi-angle and multi-dimension occurrence analysis. And through empirical research, different cases of multiple occurrence applications are selected, proving that the method can be applied to the analysis of research areas, research institutions, institutional contrast, research scholars, etc., and has good analysis results. Due to the multi-dimensional analysis and the diversification of analysis methods, this method can not only achieve the analysis effects of occurrence which includes one entity or two entities, but also reveal more extensive than the common occurrence and in-depth knowledge of content.

Keywords: multiple occurrence multiple feature items occurrence multi-source data data model knowledge discovery